DOCUMENT RESUME

ED 454 267 TM 032 876

AUTHOR Hendrickson, Amy B.
TITLE Reliability of Scores from Tests Composed of Testlets: A Comparison of Methods.
PUB DATE 2001-04-13
NOTE 37p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Seattle, WA, April 11-13, 2001).
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Reliability; Scores; Standardized Tests; Tables (Data); *Test Construction; Test Items
IDENTIFIERS *Testlets

ABSTRACT

The purpose of the study was to compare reliability estimates for a test composed of stimulus-dependent testlets as derived from item scores, testlet scores, and under the univariate generalizability theory and multivariate generalizability theory designs, as well as to determine the influence of the number of testlets and the number of items per testlet on the generalizability coefficient. For the study, random samples of 3000 examinees were drawn from the standardization data of a large standardized test. As expected, item score reliability values were largest, while reliability based on testlet scores was lowest. Generalizability coefficient estimates from the univariate and multivariate designs fell between the item and testlet reliability estimates, yet were considerably smaller (about 0.03) than the item score estimates. The multivariate analysis incorporates all items and stimulus information to obtain the most accurate reliability estimate. Four appendixes contain MGENOVA code for some item results. (Contains 10 tables and 13 references.) (Author/SLD)

Reliability of scores from tests composed of testlets:

A comparison of methods [1]

Amy B. Hendrickson [2]

University of Iowa

1. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA, April 13, 2001.
2. The author thanks Dr. Robert Brennan for comments on an earlier version of the paper.

Abstract

The purpose of the study was to compare reliability estimates for a test composed of stimulus-dependent testlets as derived from item scores, testlet scores, and under the univariate G theory p x (i:h), and multivariate G theory $p^{\bullet}$ x (i°:h°) designs, as well as to determine the influence of the number of testlets and number of items per testlet on the generalizability coefficient.

As expected, item score reliability values were largest, while reliability based on testlet scores was lowest. Generalizability coefficient estimates from the univariate and multivariate designs fell between the item and testlet reliability estimates, yet were considerably smaller (about .03) than the item score estimates. The multivariate analysis incorporates all item and stimulus information to obtain the most accurate reliability estimate.

The focus of this study is to extend previous research with the use of generalizability theory for determining the reliability of tests composed of testlets. Testlets have been described as groups of items or small tests that relate to a single content area or within which content balancing across several areas is established (Wainer & Kiely, 1987; Wainer & Lewis, 1990). Testlets may also refer to a set of items linked to a common stimulus, such as reading comprehension items relating to a passage.

There are several ways to model and scale item responses within a testlet. First, the item-stimulus relationship may be ignored all together and the items merely scored as individual units. Treating each item as an independent scoring unit, however, does not accurately reflect the measurement procedure in this case. Alternatively, stimulus information may be included in the scaling procedure by treating the item-stimulus set (or testlet) as the measurement unit.

Polytomous item response theory (IRT) models have most often been used to account for item-stimulus relationships, by modeling the item set (or testlet) as a single polytomous item. Use of polytomous IRT for testlets arose as an alternative to dichotomous IRT, whereby the item-stimulus relationship is ignored and local item independence and unidimensionality are explicit assumptions (Lord,1980). Items are locally independent if, for a given ability level, performance on one item is independent of performance on any other item. When items relate to a common stimulus performance on the items may not be independent. Thus, scoring the individual items, such as with a dichotomous IRT model, is most likely inappropriate. While using testlet units does not remove local item dependence (LID) among the items in the testlet, it allows for a way to more accurately measure performance on that set of items in relation to other test items (Yen, 1993). Due to the design and scoring of the testlet as a unit, we can be better assured of the independence of the units and the unidimensionality of the test composed of

testlets (Thissen, Steinberg, & Mooney, 1989). Thus, polytomous item response theory models have been used to account for lack of item independence by allowing for item response dependence within testlets, conditional on examinee ability, while the responses between testlets are considered to be independent. In this way, item scores are summed across each stimulus set to create a polytomous item or testlet and these testlet scores are used to score the overall test.

Testlet-based scores have been studied using polytomous IRT models to examine such characteristics as score reliability, test information, and differential item functioning (DIF; Sireci, Thissen, & Wainer, 1991; Wainer, 1995; Wainer & Lukhele, 1997). These studies found that testlet scores led to lower, but more appropriate estimates of reliability and information and could be appropriate for estimating DIF. Thus, polytomous IRT has proven useful for modeling testlets and in order to more accurately reflect the measurement procedure, compared to ignoring the item-stimulus relationship. However, the use of polytomous IRT is not without limitations.

Even though for polytomous items, these models fall under the guise of item response theory, and thus the assumptions and limitations of IRT hold. All IRT methods require that the strong statistical assumptions of unidimensionality and local independence are met (Lord, 1980). Whether at the item or testlet level, these requirements must be examined and met. Also, several polytomous IRT models exist, each with different definitions and parameterization of items. Using the various models leads to different test results. Finally, treating stimulus-related items as a testlet may lead to a loss of information from the test scores. For one, examinees with the same testlet score may not have correctly answered the same items within the testlet. Also, Yen (1993) found decreased information (increased standard errors) from testlets composed of dependent items compared to non-testlet items and to testlets composed of independent items. She suggested using testlets containing only items that show local item dependence (LID). Thus,

if six items are related to a passage but only three show LID, only those three items would be included in the testlet. She suggested that this procedure would minimize the loss of information from using testlets.

The limitations of using polytomous IRT for modeling testlets lead Lee and Frisbie (1999) to consider the use of a generalizability theory (G theory) approach to estimating reliability of scores from tests composed of testlets. Generalizability theory analysis avoids the problems of using polytomous IRT models as there is no concern about meeting strong assumptions, different scoring methods leading to different results, or loss of information. Furthermore, G theory allows for examination of the affect of including various numbers of testlets and items within each testlet on the reliability of the test scores.

Lee and Frisbie (1999) compared reliability estimates derived from three models; item-score reliability, testlet-score reliability (using the sum of item scores within each stimulus set), and a univariate G theory reliability estimate from the p x (I:H) design. Consistent with previous research and as expected, they found the item score reliability estimates to be about .04-.05 larger than both the testlet-based and G theory based estimates. They found that item-score reliability was overestimated and that the G theory approach was more appropriate for modeling the reliability of the test scores. The purpose of the present study is to replicate and extend the findings of Lee and Frisbie (1999) to include the methods of multivariate generalizability theory in assessing the reliability of stimulus-dependent item scores.

Multivariate G theory subsumes and is thus more general than univariate G theory. Both are models for identifying various sources of error from a measurement procedure. Under each model a population of objects of measurement is defined (often persons, 'p') and one or more conditions of measurement, or facets, are defined as part of a universe of admissible observations

(for example items ('i') and stimuli ('h')). Generalizability theory examines how variability in the facets affects the test scores, by partitioning out error variance due to each facet and to the interactions of the facets. For example, for a p x (i:h) design where persons are crossed with items and stimuli and items are nested within stimuli, there are five sources of error; main effects for persons (p), stimuli (h), and items within stimuli (i:h), and interaction terms for persons and stimuli (p x h), and persons and items within stimuli (p x i:h).

Generalizability analyses include generalizability studies (G studies) in which estimates of the parameter values (variance components) associated with the facets in the universe of admissible observations and the single persons in the population are calculated. Variance components are calculated for each error source from the appropriate mean squares from an analysis of variance design. For the p x (i:h) example, the following five variance components are estimated; $\hat{\sigma}^2(p)$, $\hat{\sigma}^2(i:h)$, $\hat{\sigma}^2(h)$, $\hat{\sigma}^2(ph)$, $\hat{\sigma}^2(pi:h)$.

Decision studies (D studies) are also completed for a particular universe of generalization to which the results will be generalized. These D studies involve various sample sizes of each facet and the same or different design structure as the G study. D study variance components are also calculated, by using the G study variance components and adjusting for the facet sample sizes in the universe of generalization. For example, the D study variance component for four stimuli, $\hat{\sigma}^2(H)$, is calculated as:

$$\hat{\sigma}^2(H) = \frac{\hat{\sigma}^2(h)}{4}.$$

In the D study, variance components are used to calculate the following statistics of interest:

- universe score, $\hat{\sigma}^2(p)$, similar to true score in classical test theory and like a mean score for an object of measurement (p) over all conditions in the universe of

generalization,

- absolute error variance, $\sigma^2(\Delta)$, the variance of the difference between examinee observed and universe scores,
- relative error variance, $\sigma^2(\delta)$, the same as error variance in classical test theory and the difference between examinee observed and universe scores relative to the population means for observed and universe scores, and the
- generalizability coefficient, $E\hat{\rho}^2$, a reliability-like coefficient, is calculated as

$$E\hat{\rho}^2 = \frac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p) + \hat{\sigma}^2(\delta)}.$$

In univariate G theory only one universe of generalization, defined over all facets, is of interest and any sample from this universe is considered randomly parallel to any other sample. Facets in the universe are random, such that all instances of the facet are interchangeable, or fixed, such that there are a finite number of instances of the facet defined and all instances are included in the universe of generalization. In multivariate G theory, at least one facet is fixed and one universe of generalization exists for each level of that fixed facet. The fixed facet ('v') in these designs is said to be fixed in that every form of the test involves the same categories of that facet. For example, if every test form contains a Map and a Diagram (as the test in the current study does), we can say that 'Type of Stimuli' is fixed ('v'), while the particular map or diagram used in the test is random ('h').

In multivariate G theory, the levels of the fixed facet are linked in that every person responds to all stimuli in all levels. Persons responses ('p') on the levels of the fixed facet may be correlated, therefore the design must be represented with variance/covariance matrices to account for this possibility. Thus, multivariate G theory methods involve levels of a fixed facet

with allowance for correlated scores between the levels. Other facets in the design (items, stimuli) may also be linked to the fixed facet or may be independent of the fixed facet, depending on if scores on that random facet occur at all or one level of the fixed facet. Linked and independent facets are represented in the multivariate G theory design with closed ($^{\bullet}$) and open ($^{\circ}$) circles, respectively. In the current study, the multivariate design is represented as $p^{\bullet} \times (i^{\circ}:h^{\circ})$ as persons are linked to the fixed facet and items and stimuli are nested within one level of the fixed facet. See Brennan (1992 and in press) for more complete discussions of the univariate and multivariate G theory designs.

The purpose of the current study is to:

1.) Compare reliability estimates derived from item scores, testlet scores, the univariate G theory p x (i:h), and the multivariate G theory $p^{\bullet} \times (i^{\circ}:h^{\circ})$ designs.

2.) Determine the influence of the number of testlets and number of items per testlet on the generalizability coefficients compared to item score reliability estimates.

Methods

Data

For the current study, random samples of 3000 examinees were drawn from Forms K and L from the 1992 standardization data of the Level 10 Maps and Diagrams test of the Iowa Tests of Basic Skills (ITBS; Hoover, Hieronymus, Frisbie, & Dunbar, 1994). The Form K and L Maps and Diagrams tests consist of 26 items each, distributed across two maps and two diagrams with 6, 7, and 6, 7 items each.

Analyses

Reliability of the test scores was computed in four different ways; 1.) for the 26 items (calculated as the G coefficient from a p x I design and designated as Item (α)), 2.) for the testlet-based scores (calculated as the G coefficient from a p x T design with T representing the sum of the item scores within each testlet and designated as Testlet(α)), 3.) according to a univariate p x (I:H) G theory design with H representing a random stimulus, and 4.) according to a multivariate $p^{\bullet}$ x (I°:H°) G theory design with 2 and 4 levels of the fixed facet. The two level fixed facet design represents 'Type of Stimuli' with one level being Maps and the other being Diagrams. The four level design represents combinations of 'Type of Stimuli' and 'Process Categories' from the ITBS test specifications. The test specifications list nine process categories that were combined into four categories – two corresponding to Maps and two to Diagrams. These categories were chosen to represent lower versus higher order cognitive skills and in order to have more than one item per testlet per level of the fixed facet. The categories are as follows: D1- Locate Information, Explain Relationships (with 4 items for the first Diagram and first process category and 4 items for the second Diagram and first process category for Form K and 3 and 3 items for Form L), D2 – Infer Processes or Products, Compare and Contrast Features (with

2 and 3 items for Form K and 3 and 4 items for Form L), M1 – Locate and Describe Places (with 3 and 4 items for Form K and 2 and 4 items for Form L), and M2 – Determine Distance, Interpret Data, and Infer Behavior (with 3 and 3 items for Form K and 4 and 3 items for Form L).

A G study was conducted for each design to calculate variance components for each error source. The D studies incorporated the same structures as for the G studies and produced universe scores, error variances, and G and Phi coefficients. Composite statistics (universe scores, error variances, and reliability estimates) for the multivariate G theory analyses were calculated with equal weights across the levels of the fixed facet; .5 and .5 for the two-level design; .25, .25, .25, and .25 for the four-level design. Standard errors of all calculated values were derived by using the estimates for Forms K and L, as

$$SE = \frac{|K - L|}{\sqrt{2}}.$$

Several additional D studies for the p x I, p x (I:H) and $p^{\bullet}$ x (I°:H°) designs were conducted to assess the influence of the number of testlets and the number of items per testlet on the G coefficients. In this way, the combination of numbers of items and of testlets leading to the highest reliability of the test scores could be found.

MGENOVA (Brennan, 1999) was used to run all analyses. MGENOVA is specifically designed to handle multivariate generalizability analyses, but is also able to perform the simpler univariate designs. MGENOVA uses raw scores on all persons and facets (or variance component estimates) as input, organized according to the design. The program outputs all G and D study variance and covariance components, information about the fixed facet, and statistics for estimating G and D study variance and covariance components. Also, for the D study only; sample size statistics, the universe score matrix, error matrices, and D study results for individual

variables and for the composite. Individual variance components, universe scores, error variances, and composite score results will be presented and discussed and G and Phi coefficient (reliability) estimates will be compared across the designs for various numbers of items, testlets, and items per testlet. See Appendices A-E for MGENOVA code for Item (α), Testlet (α), Univariate p x (i:h), and Multivariate 2 and 4 level $p^{\bullet}$ x (i°:h°) designs for Form K only.

Results

The G study and D study results for the univariate and multivariate generalizability analyses are presented first. Then reliability estimates from across the G theory designs as well as from classical test theory models are discussed.

Table 1 presents G and D study variance component results for the univariate p x (i:h) design for Forms K and L. Standard errors (SE) of the estimates calculated from the two forms are included in the last column. The G study variance components are fairly consistent across the two forms. Person variability is quite large compared to the other effects, though the residual pi:h terms have the highest variance components and largest SE for the two forms (SE=.0079).

The D study results in Table 1 are for the same design as the G study. Error variances are small across both forms and produce similar G and Phi coefficients. The standard error for the G coefficient is highest, indicating some variability in the reliability of scores from the two forms.

Variance and covariance estimates for the multivariate $p^{\bullet}$ x (i°:h°) design with two levels of the fixed facet representing 'Type of Stimuli' are presented in Table 2. The italicized values in the 'p' matrices show very high disattenuated correlations between examinees' performances on Maps and on Diagrams. The item ('i') and testlet ('h') effects for this design are nested in the fixed stimuli facet, so that only variances (on the diagonal) appear in the matrices for those facets. The Form K variance for Maps is considerably higher than that for Diagrams, indicating less consistency in examinees' performances on Map items compared to Diagram items on this form. However, the SE for this estimate (.0051), shows considerable variability in the estimate, itself, which would bring its value closer to the variance component estimate for Diagrams for this form. The Form L variance components for the testlet facet are much more

consistent than for Form K. The variance component estimates for the residual effects (pi:h) are, again, larger than those for the other facets.

The bottom of Table 2 shows error variance and reliability estimates for a composite of equally weighted scores (.5 and .5) from the two levels of the fixed facet. Again, error variances are small, reliability estimates (G and Phi coefficients) are relatively high, and all estimates are consistent across the forms.

Table 3 presents variance and covariance estimates for the multivariate $p^{\bullet}$ x (i°:h°) design with four levels of the fixed facet. This analysis provides more detailed information on the variability in the stimuli and in the process categories and shows consistent findings across the forms.

Table 4 summarizes differences in reliability estimates from these three generalizability theory analyses as well as those from classical test theory. As expected, the item score reliability values (Item(α)) are overestimates as indicated by the lower values from the more appropriate testlet and G theory analyses. Also, as suggested by Yen (1993), reliability based on testlet scores (Testlet(α)) appears to be an underestimate. This finding is also expected due to the lower number of 'items' used in the reliability calculation and as it was previously shown by Lee and Frisbie (1999). G coefficient estimates from the univariate p x (I:H) design fall between the item and testlet reliability estimates. This analysis allows us to incorporate all item and passage information to obtain a more accurate reliability estimate.

How do the multivariate p x (I:H) reliability estimates compare to the classical and univariate G theory coefficients? Form K and Form L G coefficients for the multivariate design are slightly larger than for the univariate p x (I:H) design, yet still considerably smaller (about .03) than the Item(α) estimates. It appears that the additional information included in the

multivariate design reveals the consistency in Diagrams across both Form K and Form L. Thus, the relative inconsistency of Maps in Form K is now 'partitioned out' and the reliability estimates increase.

To further replicate and extend the results of Lee and Frisbie (1999), several other D studies were conducted. The first set, for Form K, shown in Table 5, compare reliability estimates across designs with varying total numbers of items. As expected and previously found, reliability increases with increasing number of items and with increasing number of stimuli rather than increasing number of items per stimuli. Coefficients for the multivariate designs were affected by the pattern of number of items per type of stimuli, as would be expected from differential variability in Maps versus Diagrams items. G coefficients were higher for those designs with more or equal numbers of items per Diagram compared to the number per Map. These results are specific to this test, however, because of the differential variability in the parts of the test and as only a subset of possible designs are presented.

In the last three columns of Table 5 are the multivariate G coefficients for designs with four levels of the fixed facet ('Stimuli/Process' categories). All designs include two stimuli/process categories per level of the fixed facet, but vary in the number of items per each level of these categories. These coefficients tend to be larger than for the other designs (e.g., univariate G theory).

Table 6 summarizes the differences in the reliability estimates across these designs for Form K. Average differences are in the last row of the table and show that the largest difference in reliability estimates is between Item(α) and the multivariate design with two levels of the fixed facet. Tables 7 and 8 present the same information as Tables 5 and 6 but for Form L and show similar results as for Form K. Finally, Tables 9 and 10 show G coefficients for

multivariate $p^{\bullet}$ x (I°:H°) designs with a fixed number of items and either a fixed number of testlets (four) and a varying number of items per testlet (Table 9) or varying numbers of testlets and of items per testlet (Table 10). Table 9 shows that, for 4 testlets and 26 items, the highest reliability is achieved with the current test design, such that there are 6,7 and 6,7 items for two maps and two diagrams. Table 10 shows how the reliability estimates would vary with changes in the number of testlets and the patterns of items within these testlets. Again, these results reflect the increased variability found for Maps items (in Form K) compared to Diagrams items.

Discussion

Generally, the results from the current study show that, if appropriate to the test specifications, multivariate G theory designs may be useful in calculating an accurate estimate of the reliability of the test scores. The multivariate design incorporates more information from the test design, but also requires that additional decisions be made in using the design, such as the weighting scheme across levels of the fixed facet. This increased information better reflects the consistency or inconsistency of more aspects of the test and is incorporated in calculating the reliability of scores derived from the test.

References

Brennan, R.L. (1992). *Elements of generalizability theory* (rev. ed.) Iowa City, IA: American College Testing.

Brennan, R.L. (1999). *Manual for MGENOVA, Version 2.0.* Iowa Testing Programs Occasional Papers, No. 47. Iowa City, IA: The University of Iowa.

Brennan, R.L. (in press). *Generalizability Theory.* Springer-Verlag.

Hoover, H. D., Hieronymus, A., Frisbie, D., and Dunbar, S. (1994). *Iowa Tests of Basic Skills: Interpretive Guide for School Administrators.* Chicago, IL: The Riverside Publishing Company.

Lee, G., and Frisbie, D. A., (1999). Estimating Reliability under a Generalizability Theory Model for Test Scores composed of Testlets

Lord, F. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Sireci, S.G., Thissen, D., and Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28, 3,* 237-247.

Thissen, D., Steinberg, L., and Mooney J.A. (1989). Trace lines for testlets: A use of multiple-categorical models, *Journal of Educational Measurement, 26*(3), 247-260.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example, *Applied Measurement in Education, 8*(2), 157-186.

Wainer, H., and Kiely, G. L., (1987). Item clusters and computerized adaptive testing: A case for testlets, *Journal of Educational Measurement, 24, 3,* 185-201.

Wainer, H., and Lewis, C. (1990). Toward a psychometrics for testlets, *Journal of*

Educational Measurement, 27, 1, 1-14.

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement, 57, 5*, 741-758.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence, *Journal of Educational Measurement, 30, 3*, 187-214.

Table 1. Univariate p x (i:h) Results for Forms K and L

	Var Comp	K	L	SE[1]
G-Study	p	.0350	.0332	.0013
	h	.0008	-.0013	.0015
	i:h	.0196	.0175	.0015
	ph	.0074	.0071	.0002
	pi:h	.1826	.1938	.0079
D-Study	p	.0350	.0332	.0013
	H	.0002	-.0003	.0004
	I:H	.0008	.0007	.0001
	pH	.0019	.0018	.0000
	pI:H	.0070	.0075	.0003
	Rel Error	.0089	.0092	.0003
	Abs Error	.0098	.0096	.0002
	G-Coeff	.7976	.7824	.0107
	Phi	.7806	.7762	.0031

[1]Standard Errors Based on Estimates from Forms K and L

Table 2. Multivariate $p^{\bullet}$ x (i°:h°) Results with 2 levels of the Fixed Facet for Forms K and L

		Form K[1]		Form L[1]		SE[2]	
	VC	Diags	Maps	Diags	Maps	Diags	Maps
G-Study	p	.0309 .0346	*.9781* .0406	.0285 .0332	*1.0067* .0382	.0017 .0010	.0202 .0017
	h	-.0003	.0049	-.0024	-.0023	.0014	.0051
	I:h	.0240	.0152	.0153	.0197	.0062	.0032
	ph	.0101	.0032	.0114	.0026	.0009	.0004
	pi:h	.1834	.1818	.1967	.1909	.0094	.0065
D-Study	H	-.0002	.0025	-.0012	-.0012	.0007	.0026
	I:H	.0018	.0012	.0012	.0015	.0005	.0002
	pH	.0051	.0016	.0058	.0013	.0005	.0002
	pI:H	.0141	.0140	.0151	.0147	.0007	.0005
Compst (w =.5)	Univ Scr	.0358		.0333		.0018	
	RelError	.0087		.0092		.0004	
	AbsError	.0100		.0093		.0005	
	GCoeff	.8018		.7829		.0133	
	Phi	.7782		.7814		.0023	

[1]Italicized values are disattenuated correlations

[2]Standard errors based on estimates from Forms K and L

Table 3. Multivariate $p^{\bullet} \times (i^{\circ}:h^{\circ})$ Results with 4 levels of the Fixed Facet for Forms K and L

		Form K[1]				Form L[1]				SE[2]			
	VC	D1	D2	M1	M2	D1	D2	M1	M2	D1	D2	M1	M2
G-Study	p	.0254	*1.2271*	*1.0239*	*.9785*	.0338	*1.1344*	*.8859*	*.9394*	.0060	.0656	.0976	.0276
		.0382	.0382	*.9966*	*.9925*	.0341	.0267	*.9412*	*1.0496*	.0029	.0082	.0392	.0404
		.0314	.0374	.0369	*1.0333*	.0331	.0313	.0413	*.9504*	.0013	.0044	.0031	.0586
		.0327	.0406	.0416	.0439	.0342	.0339	.0382	.0391	.0011	.0047	.0024	.0033
	h	.0042				.0030				.0009			
			-.0010				.0302				.0220		
				.0339				.0021				.0225	
					-.0026				-.0035				.0006
	I:h	.0328				.0111				.0154			
			.0065				.0013				.0037		
				.0059				.0241				.0129	
					.0091				.0134				.0031
	ph	.0080				.0012				.0048			
			.0138				.0099				.0028		
				.0107				-.0048				.0109	
					.0008				.0089				.0058
	pi:h	.1782				.2035				.0179			
			.1930				.1970				.0028		
				.1638				.1918				.0197	
					.1974				.1876				.0069
D-Study	H	.0021				.0015				.0004			
			-.0005				.0157				.0115		
				.0173				.0011				.0115	
					-.0013				-.0017				.0003
	I:H	.0041				.0014				.0019			
			.0013				.0003				.0007		
				.0008				.0035				.0018	
					.0015				.0022				.0005
	pH	.0040				.0006				.0024			
			.0072				.0052				.0014		
				.0054				-.0024				.0056	
					.0004				.0045				.0029
	pI:H	.0223				.0254				.0022			
			.0386				.0394				.0006		
				.0234				.0274				.0028	
					.0329				.0313				.0012
Comp.s (BWTS=.25)	Universe score =				.0368				.0344				.0017
	Relative Error =				.0084				.0082				.0001
	Absolute Error =				.0100				.0097				.0002
	G-Coefficient =				.8142				.8075				.0048
	Phi =				.7866				.7802				.0045

[1]Italicized values are disattenuated correlations, [2]Standard errors based on estimates from Forms K and L

Table 4. Reliability estimates across item-score, testlet-score, p x (I:H), and $p^{\bullet}$ x (I°:H°) models

Model		Form K	Form L	Form Ave
	Item(α) (A)	.8349	.8194	.8272
	Testlet(α) (B)	.7933	.7773	.7853
U-Var	G-Coeff (C)	.7976	.7824	.7900
M-Var (2)	G-Coeff (D)	.8018	.7829	.7924
M-Var (4)	G-Coeff (E)	.8142	.8075	.8109
Differences between G-Coeff.s across Models	(A-B)	.0416	.0421	.0419
	(A-C)	.0373	.0370	.0372
	(A-D)	.0331	.0365	.0348
	(A-E)	.0207	.0119	.0163
	(B-C)	-.0043	-.0051	-.0047
	(B-D)	-.0085	-.0056	-.0071
	(B-E)	-.0209	-.0302	-.0256
	(C-D)	-.0042	-.0005	-.0024
	(C-E)	-.0166	-.0251	-.0209
	(D-E)	-.0124	-.0246	-.0185

Table 5. Form K Generalizability Coefficients of p x I, p x (I:H), and $p^{\bullet}$ x (I^{O}:H^{O}) Designs with Varying Total Number of Items

Total Number of Items	p x I Design (Item(α))		p x (I:H) Design			$p^{\bullet}$x(I^{O}:H^{O}) Design with 2 levels			$p^{\bullet}$x(I^{O}:H^{O}) Design with 4 levels		
	I'	G-Coeff (A)	H'	I"	G-Coeff (B)	H",H"'	I"	G-Coeff (C)	H's	I's	G-Coeff (D)
20	20	.7955	5	4	.7674	2,3	4	.7652	2,2,2,2	2,3;2,3;2,3;2,3	.7822
						3,2	4	.7610			
			4	5	.7612	2,2	5	.7660			
25	25	.8294	5	5	.7994	2,3	5	.7937		3,3;3,3;3,3;3,4	.8144
						3,2	5	.7991		3,4;3,3;3,3;3,3	.8138
30	30	.8537	6	5	.8270	3,3	5	.8302		5,4;3,3;5,4;3,3	.8308
			5	6	.8222	2,3	6	.8171		3,3;5,4;3,3;5,4	.8340
						3,2	6	.8228			
35	35	.8719	7	5	.8480	3,4	5	.8467		5,4;5,4;5,4;4,4	.8535
						4,3	5	.8497		4,4;5,4;5,4;5,4	.8539
			5	7	.8394	2,3	7	.8346			
						3,2	7	.8406			
40	40	.8861	8	5	.8644	4,4	5	.8670		5,5;5,5;5,5;5,5	.8674
			5	8	.8527	2,3	8	.8483			
						3,2	8	.8544			
45	45	.8975	9	5	.8776	4,5	5	.8777		6,5;6,5;6,5;6,6	.8778
						5,4	5	.8797		6,6;6,5;6,5;6,5	.8777
			5	9	.8630	2,3	9	.8593			
						3,2	9	.8655			
50	50	.9068	10	5	.8885	5,5	5	.8907		6,6;6,7;6,6;6,7	.8866
			5	10	.8721	2,3	10	.8683		6,7;6,6;6,7;6,6	.8862
						3,2	10	.8746			

Note: I'=number of items in a p x I D-Study, H' = number of passages in a p x (I:H) D-Study, I" = number of items within each passage in a p x (I:H) D-study, H" and H"' = number of passages in each level of a $p^{\bullet}$ x (I^{O}:H^{O}) D-study, and I"'= number of items within each passage of a $p^{\bullet}$ x (I^{O}:H^{O}) D-study

Table 6. Form K Differences Between G-Coeff.s for p x I, p x (I:H), and $p^{\bullet}$ x (I°:H°) Designs with Varying Total Number of Items

Total n	(A-B)	(A-C)	(A-D)	(B-C)	(B-D)	(C-D)
20	.0281	.0303	.0133	.0022	-.0148	-.0170
	.0343	.0345		.0002		
		.0295		-.0048		
25	.0300	.0357	.0150	.0057	-.0151	-.0207
		.0303	.0156	.0003	-.0145	-.0147
30	.0267	.0235	.0230	-.0032	-.0038	-.0005
	.0315	.0366	.0198	.0051	-.0117	-.0168
		.0309		-.0006		
35	.0239	.0252	.0184	.0013	-.0055	-.0068
		.0222	.0180	-.0017	-.0059	-.0042
		.0373		.0048		
		.0313		-.0012		
40	.0217	.0191	.0187	-.0026	-.0030	-.0004
	.0334	.0378		.0044		
		.0317		-.0017		
45	.0199	.0198	.0197	-.0001	-.0002	-.0001
		.0178	.0198	-.0021	-.0001	.0020
	.0345	.0382		.0037		
		.0320		-.0025		
50	.0183	.0161	.0202	-.0022	.0019	.0041
	.0347	.0385	.0206	.0038	-.0141	-.0179
		.0322		-.0025		
Ave	.0281	.0296	.0185	.0003	-.0072	-.0078

Note: A= p x I , B= px(I:H), C= $p^{\bullet}$ x (I°:H°) with two levels, D= $p^{\bullet}$ x (I°:H°) with four levels

Table 7. Form L Generalizability Coeffiecients of p x I, p x (I:H), and $p^{\bullet}$ x (I^{O}:H^{O}) Designs with Varying Total Number of Items

Total Number of Items	p x I Design (Item(α))		p x (I:H) Design			$p^{\bullet}$x(I^{O}:H^{O}) Design with 2 levels			$p^{\bullet}$x(I^{O}:H^{O}) Design with 4 levels		
	I'	G-Coeff (A)	H'	I"	G-Coeff (B)	H",H"'	I"	G-Coeff (C)	H's	I's	G-Coeff (D)
20	20	.7773	5	4	.7994	2,3	4	.7439	2,2,2,2	2,3;2,3;2,3;2,3	.7706
			4	5	.7434	3,2	4	.7385			
						2,2	5	.7456			
25	25	.8136	5	5	.7836	2,3	5	.7733		3,3;3,3;3,3;3,4	.8052
						3,2	5	.7809		3,4;3,3;3,3;3,3	.8057
30	30	.8397	6	5	.8129	3,3	5	.8134		5,4;3,3;5,4;3,3	.8265
			5	6	.8083	2,3	6	.7984		3,3;5,4;3,3;5,4	.8256
						3,2	6	.8063			
35	35	.8593	7	5	.8353	3,4	5	.8306		5,4;5,4;5,4;4,4	.8503
						4,3	5	.8350		4,4;5,4;5,4;5,4	.8500
			5	7	.8268	2,3	7	.8173			
						3,2	7	.8255			
40	40	.8747	8	5	.8528	4,4	5	.8532		5,5;5,5;5,5;5,5	.8654
			5	8	.8413	2,3	8	.8321			
						3,2	8	.8405			
45	45	.8871	9	5	.8670	4,5	5	.8645		6,5;6,5;6,5;6,6	.8771
						5,4	5	.8673		6,6;6,5;6,5;6,5	.8773
			5	9	.8530	2,3	9	.8440			
						3,2	9	.8526			
50	50	.8972	10	5	.8787	5,5	5	.8790		6,6;6,7;6,6;6,7	.8869
			5	10	.8625	2,3	10	.8538		6,7;6,6;6,7;6,6	.8871
						3,2	10	.8624			

Note: I'=number of items in a p x I D-Study, H' = number of passages in a p x (I:H) D-Study, I" = number of items within each passage in a p x (I:H) D-study, H" and H"' = number of passages in each level of a $p^{\bullet}$ x (I^{O}:H^{O}) D-study, and I"'= number of items within each passage of a $p^{\bullet}$ x (I^{O}:H^{O}) D-study.

Table 8. Form L Differences Between G-Coeff.s for p x I, p x (I:H), and $p^{\bullet}$ x (I°:H°) Designs with Varying Total Number of Items

Total n	(A-B)	(A-C)	(A-D)	(B-C)	(B-D)	(C-D)
20	-.0221	.0334	.0067	.0555	.0288	-.0267
	.0339	.0388		.0049		
		.0317		-.0022		
25	.0300	.0403	.0084	.0103	-.0216	-.0319
		.0327	.0079	.0027	-.0221	-.0248
30	.0268	.0263	.0133	-.0005	-.0136	-.0131
	.0314	.0413	.0142	.0099	-.0173	-.0272
		.0334		.0020		
35	.0240	.0287	.0090	.0047	-.0150	-.0197
		.0243	.0093	.0003	-.0147	-.0150
		.0420		.0095		
		.0338		.0013		
40	.0219	.0215	.0093	-.0004	-.0126	-.0122
	.0334	.0426		.0092		
		.0342		.0008		
45	.0201	.0226	.0100	.0025	-.0101	-.0126
		.0198	.0098	-.0003	-.0103	-.0100
	.0341	.0431		.0090		
		.0345		.0004		
50	.0185	.0182	.0103	-.0003	-.0082	-.0079
	.0347	.0434	.0101	.0087	-.0246	-.0333
		.0348		.0001		
Ave	.0239	.0328	.0098	.0058	-.0118	-.0195

Note: A= p x I , B= px(I:H), C= $p^{\bullet}$ x (I°:H°) with two levels, D= $p^{\bullet}$ x (I°:H°) with four levels

Table 9. Composite[1] Generalizability Coefficients of the $p^{\bullet}$ x (I^{O}:H^{O}) Design with Fixed Total Number of Items and Fixed Number of Stimuli and Varying Number of Items within Each Stimuli.

Total n	Fixed H' (2,2) Varying I'	K G-coeff	L G-coeff	SE
26	4,9;4,9	.7975	.7784	.0135
	3,10;3,10	.7932	.7739	.0137
	4,10;5,7	.7968	.7775	.0137
	5,7;4,10	.7992	.7803	.0133
	5,9;6,6	.7994	.7803	.0135
	6,6;5,9	.8006	.7817	.0134
	6,7;6,7	.8018	.7829	.0133

[1] Weights =.5, .5

Table 10. Composite[1] Generalizability Coefficients of the $p^{\bullet}$ x (I°:H°) Design with Fixed Total Number of Items and Varying Number of Stimuli and Varying Number of Items within Each Stimuli.

Total n	Total H	H'	I'	K	L	SE
26	2	1,1	13,13	.7727	.7521	.0146
	3	2,1	8,9;9	.7819	.7639	.0127
		1,2	9;8,9	.7668	.7437	.0163
	4	2,2	6,7;6,7	.8018	.7829	.0133
	5	3,2	5,5,5;5,6	.8066	.7889	.0125
		2,3	6,5;5,5,5	.8010	.7812	.0140
	6	3,3	4,4,5;4,4,5	.8120	.7938	.0129
	7	3,4	4,4,4;4,4,3,3	.8126	.7940	.0131
		4,3	3,3,4,4;4,4,4	.8152	.7976	.0124
	8	4,4	3,3,3,4;3,3,3,4	.8172	.7993	.0126
	9	5,4	2,3,3,3,3;3,3,3,3	.8189	.8015	.0124
		4,5	3,3,3,3;3,3,3,3,2	.8173	.7992	.0128
	10	5,5	2,2,3,3,3;2,2,3,3,3	.8202	.8025	.0125

[1] Weights =.5, .5

APPENDICES

APPENDIX A

MGENOVA code for Item(α) results - Form K

```
GSTUDY   M&D p x i Design  Grade 4 K
OPTIONS  NREC 4  "*.out" EMS ET DEFAULT_DSTUDY
MULT     1 Stimuli
EFFECT   * p  3000
EFFECT   # i  26
FORMAT   0 0
PROCESS  "Grade4K"
DSTUDY   M&D p x I Sample Size Differ
DEFFECT  $ p  3000
DEFFECT  # I  13
ENDDSTUDY
DSTUDY   M&D p x I Sample Size Differ
DEFFECT  $ p  3000
DEFFECT  # I  20
ENDDSTUDY
DSTUDY   M&D p x I Sample Size Differ
DEFFECT  $ p  3000
DEFFECT  # I  25
ENDDSTUDY
DSTUDY   M&D p x I Sample Size Differ
DEFFECT  $ p  3000
DEFFECT  # I  30
ENDDSTUDY
DSTUDY   M&D p x I Sample Size Differ
DEFFECT  $ p  3000
DEFFECT  # I  35
ENDDSTUDY
```

APPENDIX B

MGENOVA code for Testlet (α) results – Form K

```
GSTUDY   M&D p x t Design  Grade 4 K
OPTIONS  NREC 4  "*.out" EMS.ET DEFAULT_DSTUDY
MULT     1 Stimuli
EFFECT   * p  3000
EFFECT   # t  4
FORMAT   0 0
PROCESS  "Grade4K"
DSTUDY   M&D p x I Sample Size Differ
DEFFECT  $ p  3000
DEFFECT  # T  4
ENDDSTUDY
```

APPENDIX C

Selected MGENOVA code for univariate G theory p x i:h results – Form K

```
GSTUDY   M&D p x (i:h) Design  Grade 4 K
OPTIONS  NREC 4  "*.out" EMS ET DEFAULT_DSTUDY
MULT     1 Stimuli
EFFECT   * p    3000
EFFECT   # h    4
EFFECT   # i:h  6 6 7 7
FORMAT   0 0
PROCESS  "Grade4K"
DSTUDY   M&D p x (I:H) Sample Size Differ
DEFFECT  $ p    3000
DEFFECT  # H    4
DEFFECT  # I:H  3 3 4 4
ENDDSTUDY
DSTUDY   M&D p x (I:H) Sample Size Differ
DEFFECT  $ p    3000
DEFFECT  # H    2
DEFFECT  # I:H  13 13
ENDDSTUDY
DSTUDY   M&D p x (I:H) Sample Size Differ
DEFFECT  $ p    3000
DEFFECT  # H    3
DEFFECT  # I:H  8 9 9
ENDDSTUDY
DSTUDY   M&D p x (I:H) Sample Size Differ
DEFFECT  $ p    3000
DEFFECT  # H    5
DEFFECT  # I:H  5 5 5 5 6
ENDDSTUDY
```

APPENDIX D

Selected MGENOVA code for multivariate G theory p x i:h results – Form K

```
GSTUDY    M&D p x (i:h) Mulitvariate Design  Grade 4 K
OPTIONS   NREC 4  "*.out" EMS ET DEFAULT_DSTUDY
MULT      2 Diagrams Maps
EFFECT    * p    3000 3000
EFFECT     h    2 2
EFFECT     i:h  6 7
                  6 7
FORMAT    0 0
PROCESS   "Grade4KM"
DSTUDY    M&D p x (I:H) Sample Size Differ
DEFFECT   $ p    3000 3000
DEFFECT    H    2 2
DEFFECT    I:H  3 3
                  4 4
ENDDSTUDY
DSTUDY    M&D p x (I:H) Sample Size Differ
DEFFECT   $ p    3000 3000
DEFFECT    H    1 1
DEFFECT    I:H  13
                  13
ENDDSTUDY
DSTUDY    M&D p x (I:H) Sample Size Differ
DEFFECT   $ p    3000 3000
DEFFECT    H    2 1
DEFFECT    I:H  8 9
                  9
ENDDSTUDY
```

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM032876

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Reliability of scores from tests composed of testlets: A comparison of methods	
Author(s): Amy B Hendrickson	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1 ↑

[X]

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A ↑

[ ]

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B ↑

[ ]

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Amy B. Hendrickson	Printed Name/Position/Title: Amy B. Hendrickson	
Organization/Address: University of Iowa, 332 LC, Iowa City, IA. 52246	Telephone: 319-335-5419	FAX: 319-335-6038
	E-Mail Address: amy-b-hendrickson@uiowa.edu	Date: 5/7/01

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com